COMP47750

Machine Learning with Python

# Autumn 2024 Assignment 1

# **Algorithm Bias**

## 

## Objective

The objective of this assignment is to assess the impact of algorithm bias on machine learning algorithms trained with imbalanced data. Submissions should also propose and evaluate strategies for overcoming this bias. The dataset to be used in this assignment is provided in the csv file **appendicitis.csv.** This dataset is derived from the UCI dataset here: <https://archive.ics.uci.edu/dataset/938/regensburg+pediatric+appendicitis>.

## Requirements

1. Load the Appendicitis dataset into a Python notebook and assess the bias of *k-*NN and Decision Tree classifiers trained on this dataset, i.e. are classifiers biassed towards the majority class? Be careful to select a meaningful metric to quantify bias and ensure that the estimates from your testing are as low-variance as possible given the size of the dataset.
2. Working with the Decision Tree classifier, propose one or more strategies to rectify any bias with that classifier. You can use strategies from the Imbalanced-learn library (see link below), you can use strategies available within scikit-learn or you can propose and implement your own. Comment on any improvements achieved.
3. Repeat the exercise in Task 2 for *k-*NN. You may find that the strategies that work in Task 2 will not work here. Comment on any improvements achieved and compare with the results in Task 2.

You should use plots (matplotlib) and commentary in markdown to report your results.

**Hints**:

**Imbalanced-learn** is a Python library compatible with scikit-learn for dealing with imbalanced data: <https://imbalanced-learn.org/stable/index.html>

**Submission:** This is an individual (not group) project. Submission is through the Brightspace page. Your submission should comprise your original notebook as an **ipynb** file and also saved as **html**. If the notebook takes a while to run you should submit it without clearing the outputs.